AD-A237 856

‖‖‖‖‖‖‖‖‖‖‖‖

~~REPORT DOCU~~MENTATION PAGE

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
		FINAL TECH 01 Aug 88 to 31 Jul 90

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
ANALOG COMPUTATION IN NEURAL SYSTEMS: ARCHITECTURES AND COMPLEXITY	AFOSR-88-0227 61102F 2304/A2

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) PRINCETON UNIVERSITY DEPT OF ELECTRICAL ENGINEERING PRINCETON, NEW JERSEY 08544-5263	8. PERFORMING ORGANIZATION REPORT NUMBER AFOSR·TR· 91 0554

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR/NM Bldg 410 Bolling AFB DC 20332-6448	10. SPONSORING/MONITORING AGENCY REPORT NUMBER AFOSR-88-0227

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited.	12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The research efforts supported by AFOSR Grant AFOSR-88-0227 concentrated on three sorts of neural network problems. First, we studied the representation problem for the class of single-hidden-layer feedforward networks, which is fundamental for understanding limitations of learning algorithms, and which also contributed to understanding the behavior of learnig algorithms in applications involving low-complexity networks. The second kind of problem studied concerns dynamics behavior in neural networks containing feedback (trellis-structured networks in one particular applications). Our work focused on studying stability issues and exploring the implications of computational complexity theory. Third, the PAC learning paradigm (probably Almost Correct) was analyzed with the goal of characterizing the effects of statistically dependent sequences of training examples on learning performance. The goal of all these efforts was to discover and explore insights about fundamental limitations on the computational capabilities of analog neural systems and, where possible, of more general classes of physical systems as well.

14. SUBJECT TERMS			15. NUMBER OF PAGES
			16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)

PRINCETON UNIVERSITY
Department of Electrical Engineering
Princeton, New Jersey 08544-5263

Final Technical Report: Period 1 Aug. 1988 to 31 July 1990

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH GRANT AFOSR-88-0227

Analog Computation in Neural Systems: Architectures and Complexity

Program Officer: Dr. M. Jacobs, AFOSR/NM

Bradley W. Dickinson
Professor
Principal Investigator
(609)-258-2916/4644

Report date: May 17, 1991

A-1

91   10   024

Grant AFOSR-88-0227 — Final Technical Report

The research efforts supported by AFOSR Grant AFOSR-88-0227 concentrated on three sorts of neural network problems. First, we studied the representation problem for the class of single-hidden-layer feedforward networks, which is fundamental for understanding limitations of learning algorithms, and which also contributes to understanding the behavior of learning algorithms in applications involving low-complexity networks. The second kind of problem studied concerns dynamic behavior in neural networks containing feedback (trellis-structured networks in one particular application). Our work focused on studying stability issues and exploring the implications of computational complexity theory. Third, the PAC learning paradigm (Probably Almost Correct) was analyzed with the goal of characterizing the effects of statistically dependent sequences of training examples on learning performance. The goal of all of these efforts was to discover and explore insights about fundamental limitations on the computational capabilities of analog neural systems and, where possible, of more general classes of physical systems as well.

We turn first to a discussion of our work on the representation problem. Despite its importance and, as it turns out, its amenability to study with familiar tools from functional analysis, the characterization of functions that may be represented by single-hidden-layer feedforward (SHLFF) neural networks has only very recently been rigorously determined. That is not to say that the answer was unexpected! Because of extensive empirical evidence, reinforced by powerful mathematical results like Kolmogorov's representation for functions of several variables, it was reasonable to believe that the class of functions implemented by SHLFF neural networks was at least a "large" subset of some class of well-behaved functions. Using the Hahn-Banach theorem on a suitable formalization of this problem, George Cybenko gave an elegant proof that the SHLFF neural networks generate a dense set of functions in suitable topologies; other similar results have been obtained by Funahashi, by Hornik, Stinchcombe, and White, and by Jones and Barron. The result holds for a wide class of neuron (sigmoidal) nonlinearities, but we wish to emphasize that once a particular choice of nonlinearity has been made, each node of the network uses it (which is in distinction to results which follow from Kolmogorov's representation). Thus the functions which are dense

are finite linear combinations of the form

$$F_{SHLFF}(x_1, \ldots, x_n) = \sum_{i=1}^{N} a_i \sigma(\sum_{k=1}^{n} b_k x_k)$$

The "price" of elegance in Cybenko's approach (and in all the others mentioned except for the very recent work of Jones and Barron) is its existential, rather than constructive, solution to the problem of approximation. Our work has attacked the approximation problem directly using techniques related to the Radon transform and reconstruction from projections. (These tools are familiar ones in the signal processing literature dealing with tomography, where functions of 2 and 3 spatial variables arise.) This allows us to develop error bounds which relate the number of nodes in the required hidden layer, N in the above equation, to smoothness properties of the function being approximated and to the quality of the desired approximation. Furthermore, the calculations required to carry out the approximation are standard ones from signal processing; an approximation for the "toy" XOR function can even be carried out analytically. A paper describing the constructive approximation approach to representation by SHLFF neural networks was presented at the International Joint Conference on Neural Networks in Washington, June 1989 [1]. In view of the recent work of Jones and Barron, it would be of interest to investigate how the error bound in [1] might be sharpened since at least under some assumptions about regularity of the underlying function, approximations accurate to order $1/N_h$, i.e. accuracy inversely proportional to the number of hidden nodes, can be obtained.

Our work involving extensions of the constructive approximation have concerned a set of more "practical" issues related to the numerical conditioning of approximations methods based on the SHLFF architecture. In this later work [2], the analysis of the underlying one-dimensional approximation problems was considered. Some interesting conclusions were reached. First, the inherent advantage of the sigmoidal nonlinearity used in neural networks over classical polynomial nonlinearities is clearly revealed. The inability of polynomial interpolating functions to give uniformly good function approximations (indeed, the sup norm of the approximation error can grow exponentially fast with the number of interpolating points!) is well known in approximation theory. This is one of the strongest motivations for

the use of interpolating splines for function approximation. With splines, the sup norm of the approximation error decreases as more interpolating points are added; the rate of improvement in approximation depends on the smoothness of the underlying function. Furthermore, linear and cubic splines can be uniformly approximated with a logistic sigmoidal function of the kind commonly used in neural networks, and thus the sigmoidal neural networks inherit the excellent approximation capabilities that are associated with splines.

A second conclusion duplicates one obtained earlier in [1]. A fixed set of input/hidden layer connection weights may be preselected (once the number of hidden nodes is chosen and the approximation error bound is fixed) without sacrificing the "universal approximation" capabilities of the network. In terms of the analogy with spline approximation, this corresponds to the fact that good approximation can be achieved with preselected interpolation points (knots). Consequently, the SHLFF network may basically take the form of a CMAC network, where input/hidden layer structure and weights remain fixed while hidden/output layer weights are adjusted (perhaps through a training procedure or other learning mechanism) in order to give a good fit based on sample values of the function. In the case of mean square error approximation, the main advantage is that no "error backpropagation" is required for training, and this means that the amount of computation required for training can be drastically reduced.

A third significant result is an explicit motivation for the use of paired sigmoidal functions, in the form of a difference. $\sigma(x-L) - \sigma(x+L)$, rather than independent sigmoidal functions, based on considerations of numerical conditioning of mean square error approximation. While the condition number for approximation with sigmoidal functions grows only linearly with the number of hidden nodes N_h, a constant condition number can be achieved using paired sigmoids because the difference of two shifted sigmoids is effectively of finite support. This intrinsic difficulty with scaling behavior of training algorithms for SHLFF has not been previously made explicit such algorithms should be expected to require a logarithmically growing word length to maintain enough accuracy to obtain valid solutions. It is to be expected that the paired sigmoidal networks display much improved scaling behavior. and the may be useful for empirical studies of the scaling of intrinsic difficulty of classes of

problems for which backpropagation training is routinely applied.

Dynamical behavior of neural networks involving feedback interconnections has attracted attention for a variety of applications: (Hopfield) associative memory, optimization, and efficient realization of competitive selection processes such as winner-take-all networks. Studying analog neural networks amounts to applying such well-known ideas as Lyapunov functions, gradient flows, etc,. from differential equations and dynamical systems. Of some interest is the limiting behavior of a parametrized family of analog neural networks which are intended to approach a suitable discrete model, e.g one employing binary threshold elements rather than sigmoidal nonlinearities. Arguments about how analog behavior approaches some kind of discrete behavior in the limit of infinite gain sigmoidal nonlinearities have been advanced and experimentally validated, typically using a sequence of "frozen" gain (piecewise constant gain) networks.

Our work focused mainly on an interesting stability problem for trellis networks involving a sequence of identical layers with a fixed pattern of feedforward connections between layers and with "on-center, off-surround" competitive feedback interconnections within each layer. Such networks were developed some time ago using maximum likelihood sequence estimation for convolutional decoding as a motivation, and they have been shown to provide error-correcting capabilities while allowing for an efficient means of incorporating fault tolerance [3]. Besides an ongoing theoretical investigation of the stability issues involving trellis networks, we have worked on exploring how this architecture may be extended to obtain the capacity for "self-repair" in networks; our approach is to introduce one or more redundant spare neurons at each stage of the trellis. It turns out that the self-repairing process can proceed at the same time gradient-type training is being used. The automatic replacement of faulty neurons by spares is accomplished at some loss in the speed of convergence of learning as would be expected. A paper discussing this work has appeared in the *IEEE Trans. on Neural Networks* [3], and another survey of this general set of problems was presented at the 1990 Conference on Decision and Control [4].

As an attempt to exploit extra hardware for increasing the learning rate of feedforward networks, we explored analogies with cooperating parallel systems occurring in nature. In particular, a model inspired by the flocking behavior of birds and the schooling behavior of

fish has been formulated to investigate loosely-coupled parallel learning processes for SHLFF networks according to a back-propagation paradigm. A paper on this subject was presented at the 1989 Conference on Information Sciences and Systems at Johns Hopkins [5], and another paper was presented as a contribution to an invited session on Neural Networks for the 1989 IEEE Conference on Decision and Control in Tampa [6]. A fully distributed version of the parallel learning algorithm, together with constraints that insure (asymptotically) that learning is not accomplished at the expense of good generalization capabilities, remain for further investigation.

We also pursued our idea of connecting some notions from computational complexity theory with limitations on performance of analog neural networks (as the particular case of analog computing system of primary interest). Here our work was very much exploratory, and we looked into neural network models where chaotic dynamics might play a necessary role in allowing the computation of solutions corresponding to intractable, i.e. NP-complete, combinatorial optimization problems. We also explored how even simpler combinatorial problems can lead to problematical analog neural network "solution" methods. For linearly-separable binary classification problems, using work by Sontag, we showed that the discretization of continuous-time gradient learning algorithms with finite "margins," which converge in finite time, produces a polynomial-time algorithm. Some very simple polynomial-time algorithms for trivial combinatorial problems such as matrix transposition using Hopfield-type networks were obtained. A talk on our work relating computational complexity and neural network behavior was presented as part of an invited session at the 1988 IEEE Conference on Decision and Control in Austin [7]. Overall, we are not able to prove any general results that would bear on the use of continuous-time neural networks for solving combinatorial optimization problems. Perhaps it is the case that only combinatorial problems solvable in linear-time can be solved with such a form of analog computation (i.e. with differential equations such as those describing a continuous-time neural network). Brockett's analog sorter, described by gradient dynamics on a manifold of orthogonal matrices, is an example showing that $n\log n$ combinatorial problems need not have fast (i.e. polynomial time) analog solutions.

To tackle learning problems within a satisfying mathematical framework, we studied

some important work on learning theory by Valiant (the PAC, or probably almost correct, learning paradigm), and on performance bounds for multilayer perceptrons by Baum and others where the useful notion of VC-dimension provides a means of characterizing performance limitations. This work has suggested to us the need to investigate models where examples are not drawn independently, since this is the case of practical interest in most signal processing applications. Baum himself has looked into learning from example and queries, and the queries may be thought of as an extreme (highly beneficial) form of dependence of examples. Our work dealt with a more "classical" statistical setting, motivated by the process of sampling from a population without replacement.

Our method was to recast the known results on sample complexity for independent sampling to explicitly account for dependence. A new version of the proof showing how sample size grows with the VC-dimension of the class of candidate hypotheses in Valiant's PAC learning model was developed. The proof displays the properties of certain conditional probabilities that assure that learning is as least as fast as for independent sampling. Basically, the conditional probabilities must promote the sampling of a sufficiently rich set of examples, and do so at least as well as examples generated by independent sampling. For the case of examples generated as samples of low-dimensional Markov chains, an augmented state model related to first return times may be used to compute the conditional probabilities that must be checked to verify improved learning rates. Considerably more work is necessary to gain a full understanding of many issues. It appears that a learning theory based on robust statistical inference rather than nonparametric inference may be better able to provide answers to practical questions about learning from dependent examples, and this remains as a topic for future research.

Publications supported through grant AFOSR-84-0381

A list of publications prepared with the support of the grant follows, numbered according to the citations in the preceding report of accomplishments. In addition, much of the work described in the Ph.D. thesis of Dr. Sean M. Carroll, now an Assistant Professor in the Department of Electrical Engineering at Tri-State University, Angola, Indiana, was supported by the grant. The thesis, entitled *Intelligent Least Squares Methods*, was completed in August, 1990. A copy of the Abstract of Dr. Carroll's dissertation is attached to this report.

References

1. S.M. Carroll and B.W. Dickinson, "Construction of neural nets using the Radon transform," *Proceedings, Internat. Joint Conf. on Neural Networks*, Washington, DC. 1989, pp. I.607 - I.611.

2. S.M. Carroll and B.W. Dickinson, "Approximating functions using neural nets: Analysis of the scalar approximation problem," submitted for publication.

3. T. Petsche and B.W. Dickinson, "Trellis codes, receptive fields, and fault tolerant, self-repairing neural networks," *IEEE Trans. on Neural Networks*, vol. 1, 1990, pp. 154-166.

4. B.W. Dickinson, "Structured neural networks for fault tolerant performance," *Proceedings, 29th Conf. on Decision and Control*, Honolulu, HI. 1990, pp. 2741-2743.

5. B.W. Dickinson, "Learning in structured, layered neural networks," *Proceedings, 1989 Conf. on Information Sciences and Systems*, The Johns Hopkins Univ., Baltimore, MD. pp. 594-596.

6. B.W. Dickinson, "Group behavior models for learning in neural networks," *Proceedings, 28th Conf. on Decision and Control*, Tampa, FL., Dec. 1989, pp. 249-251.

7. B.W. Dickinson, "Analog neural systems" (Summary), *Proc. 27th IEEE Conf. on Decision and Control*, Austin, TX. Dec. 1988, p. 798.

8. S.M. Carroll and B.W. Dickinson, "Learning from correlated examples," under revision and to be submitted for publication.

Intelligent Nonlinear Least-Squares Methods

Sean Michael Carroll

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE
BY THE DEPARTMENT OF
ELECTRICAL ENGINEERING

June 1990

Intelligent Nonlinear Least-Squares Methods

Sean Michael Carroll

Abstract

This thesis investigates some mathematical problems arising in the modeling of input-output mappings. The investigation is motivated by the need to confront directly the nonlinearities that arise in the construction of approximate models. After a brief introduction in Chapter 1, Chapter 2 presents a model-simplification algorithm for discrete-time linear systems by assuming that certain operators commute, then invoking linear least-squares methods to produce near-optimal solutions. Chapters 3 and 4 investigate function approximation by means of neural networks, interconnected systems of simple elements which are nonlinear and adaptive. Chapter 3 develops an analytical, constructive approximation procedure applicable to a large class of functions. This is accomplished by exploiting an analogy between an operator involved in the inverse Radon Transform and the operation performed by a feedforward neural network with a single hidden layer. The high-dimensional approximation problem is reduced to a series of scalar approximation problems involving projections of the original function onto lines. Chapter 4 describes the efficient solution of such scalar approximation problems by noting the generally good performance of B-splines and

demonstrating that neural networks can be structured to behave in very similar ways Chapter 5 studies the problem of learning concepts from positive and negative examples. When the examples are chosen randomly and independently, it is known how many examples are needed; our work extends this result to the more general case of correlated examples. The results are interpreted to show that the statistical behavior agrees with intuition. We also formulate an alternative model of learning based on statistical robustness rather than on the nonparametric theory already in use. A concluding chapter reviews the techniques used and open problems uncovered in the previous chapters.